基于字簇的多模型中文分词方法研究 *

李对红，王裴岩†，张桂平，张少阳

(沈阳航空航天大学 人机智能研究中心，沈阳 110136)

摘 要：字标注分词方法是当前中文分词领域中一种较为有效的分词方法。但由于中文汉字本身带有语义信息，不同字在不同语境中其含义与作用不同，导致每个字的构词规律存在差异。针对这一问题，提出了一种基于字簇的多模型中文分词方法。该方法首先对每个字进行建模，然后对学习出的模型参数进行聚类分析形成字簇，最后基于字簇重新训练模型参数。实验结果表明，该方法能够有效地发现具有相同或相近构词规律的字簇，很好地区别了同类特征对不同字的作用程度。

关键词：中文分词；构词规律；模型参数；聚类

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2018.08.0540

Multi-model Chinese word segmentation method based on character clusters

Li Duihong, Wang Peiyan†, Zhang Guiping, Zhang Shaoyang

(*Human-Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang* 110136, *China*)

Abstract: Character-based tagging method is currently an effective method in Chinese word segmentation. However, the Chinese characters had their own semantic information, different characters had different meanings and functions in different contexts, which lead to different correlations with context, resulting in the difference of word-formation rules for each word. To solve this problem, this paper proposed a multi-model method based on character cluster. Firstly, the method separately constructed a model for each word, then clustered the model parameters to form character clusters , and finally retrained the model parameters based on the character clusters. Experimental results show that this method can effectively find character clusters with the same or similar word-formation rules, and distinguish the effect of similar features for different characters.

Key words: Chinese word segmentation; word-formation rules; model parameters; clustering

0 引言

词是能够独立运用的最小语言单元。与英语和其他西方语言有所不同，中文以字为基本书写单位，词语之间没有明显的分界符加以区分，如果不进行分词，计算机就无法得知中文词的确切边界，从而很难理解文本中所包含的语义信息[1]。因此，中文分词是自然语言处理中的一项基础性工作，其在命名实体识别、文本自动分类、机器翻译等领域都有着举足轻重的地位，其性能的好坏直接影响后续的自然语言处理任务。

中文分词方法中，有指导的字标注分词方法[2]具有较好的分词效果。该方法将分词过程抽象为序列标注任务，采用适合于序列标注的机器学习模型进行建模。其中，应用比较广泛的序列标注模型主要有最大熵马尔可夫模型(maximum entropy Markov model，MEMM)[3]、隐马尔可夫模型(hidden Markov model，HMM)[4]和条件随机场(conditional random field，CRF)模型[5-7]。然而，这些模型分词效果的好坏很大程度上受制于特征的选择和提取。近年来，随着深度学习的蓬勃发展，循环神经网络(recurrent neural networks，RNN)、长短期记忆(long-short term memory，LSTM)神经网络以及它们的变体等适用于序列标注任务的神经网络模型被广泛地应用于分词任务[8~11]。

无论是传统的机器学习模型还是神经网络模型，它们都是着眼于句子中的每一个字(或是符号)，根据当前待标注字的上下文环境判断其词位信息，以此作为分词的标记。该类方法基于训练语料建立单一模型参数，考虑的是上下文环境对所有字的全局综合作用，即假设相同上下文环境对不同待标注字的影响相同，学习出字构词的一般性规律。然而，由于中文汉字本身带有语义信息，造成了每个字的构词规律存在差异，即使相同的字作为待标注字的上下文特征时其含义与作用也存在较大的差异[12,13]，造成与待标注字的结合紧密程度发生变化。以如下例子进行说明：

a) 建立/稳定/和睦/的/两岸/关系/。/

b) 营造/了/民主/和谐/的/气氛/。/

c) 党中央/坚持/领导/和/党/的/十五大/精神/。/

上述的三个例句，当前待标注字分别为“睦”、“谐”、“党”时，前一个特征都为“和”，然而，相同的特征对待标注字的影响却不同，即与待标注字的结合紧密程度不同。从例子中可以看出，一、二句中的“和”与待标注字“睦”、“谐”的结合紧密程度相同，而第三句中的“和”与待标注字“党”的结合紧密程度与前两例句不同。因此，假设上下文环境对待标注字的影响相同显然存在问题。针对这一问题，文献[14]

收稿日期：2018-08-06；**修回日期：**2018-10-04 **基金项目：**辽宁省自然科学基金计划重点项目（20170540705）；国家教育部人文社会青年科学研究基金资助项目（17YJC740087）

作者简介：李对红（1992-），女，辽宁朝阳人，硕士研究生，主要研究方向为人工智能与自然语言处理；王裴岩（1983-），男（通信作者），讲师，博士，主要研究方向为机器学习、信息抽取（wangpy@sau.edu.cn）；张桂平（1962-），女，教授，博士，主要研究方向为自然语言处理与机器翻译、知识工程与知识管理；张少阳（1991-），男，辽宁鞍山市人，硕士，主要研究方向为人工智能与自然语言处理.

提出了基于字的多模型中文分词方法，该方法最大的特点是对每个字建立单独的模型参数，有效地区分了相同特征对不同待标注字的影响，学习出了字构词的特殊性规律。然而，该方法也存在不足，尽管针对每个字训练得到的模型参数可以很好地反映该字的构词规律，但是，存在某些字的构词规律相同或相近，势必造成模型参数的冗余。并且部分字的训练样本较少，也会造成未登录词召回率的降低。

本文对基于字的多模型中文分词方法[14]进行改进，提出了一种基于字簇的多模型分词方法。该方法对基于字的多模型分词方法学习出的模型参数进行聚类分析，将具有相同或相近构词规律的字聚合形成字簇，并基于此字簇训练模型参数。该方法与单模型方法相比，有效地提高了词表词召回率，与多模型方法相比，有效地提高了未登录词召回率，并在 PKU 语料与 MSR 语料上得到了验证。

1　基于字簇的多模型分词方法

1.1　模型训练流程

本文提出了一种基于字簇的多模型分词方法，该方法首先基于字的多模型分词方法训练得到每个字的模型参数。字的模型参数代表了该字的构词规律。接下来对上述的模型参数进行聚类分析，发现模型参数之间内在的分布结构，将具有相同或相近构词规律的字聚合形成字的类簇。最后，基于上述的类簇重新训练得到模型参数。其具体训练流程包括三个部分：字模型参数获取、字构词规律分布结构发现、模型再训练。如图 1 所示。

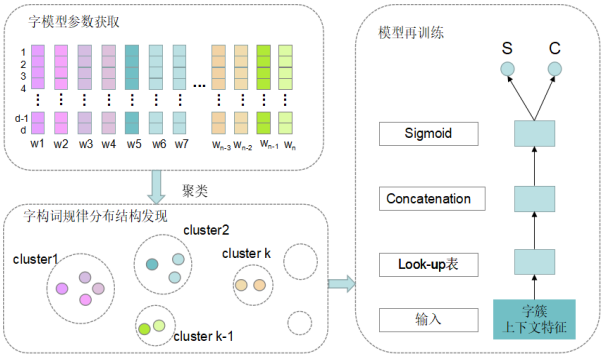

图 1　基于字簇的多模型分词方法模型训练流程

Fig. 1　The training process of multi-model segmentation method based on character clusters

1)字模型参数获取

本文基于字的多模型分词方法训练得到模型参数，各个模型参数之间相互独立，有效地学习出每个字的构词规律。图 1 显示了该方法训练得到的 n 个字的模型参数，每个模型参数由 d 维向量表示，代表了该字的构词规律。

2)字构词规律分布结构发现

本文所提方法的关键之处在于如何发现上述模型参数之间的相关性，学习多个字之间共有的构词规律。本文采用层次聚类算法对上述得到的模型参数进行聚类分析，形成字的类簇，其中每一类簇表示该类字具有相同或相似的构词规律。具体另见 1.4 小节。

3)模型再训练

该模块根据上述生成的字的类簇重新在训练语料中抽取训练样本，输入到模型结构中进行训练。具体模型结构另见 1.2 小节。

1.2　模型结构

中文分词过程通常被视为字符级别的序列标注问题，因此，可以将分词过程视为对字符串中的每个字符标注的机器学习过程。本文借鉴 Ma Jianqiang 等人[15]的思想，将模型结构分为 3 个部分，分别为 Look-up 表、Concatenation 函数、Sigmoid 函数。与基于字的多模型分词方法有所不同，本文的分词方法基于字簇进行建模。分词时，根据对应的字簇模型参数进行决策。为了降低问题的复杂性，每个类簇模型采用相同的结构。具体模型结构如图 2 所示。

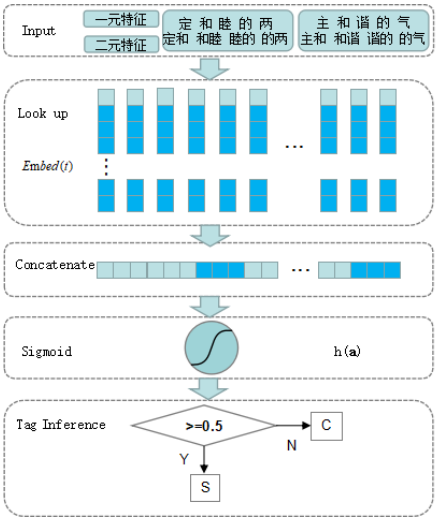

图 2　模型结构

Fig. 2　Model structure

a)Look-up 表,记录了特征与实数向量之间的映射关系，又称为特征的 Embedding。每个不同特征 t 的 Embedding 记 $Embed(t)=\mathbb{R}^N$，其中 N 表示实数向量的维度。特征是从训练语料中提取的。

b)Concatenation 函数。为了预测待标注字符的标记状态，需要将其对应的特征 Embedding 连接成一个单一的向量，作为模型的输入，记为 $a\in\mathbb{R}^{N\times K}$，其中 K 是用于描述待标注字符的特征数量，N 为特征 Embedding 的维度。

c)Sigmoid 函数。模型结构中采用的激活函数为 Sigmoid 函数，定义如式(1)所示。其中，a 为输入的特征 Embedding，w 为特征权重，$\langle a,w\rangle$ 表示两个向量的点积。

$$h(a)=\frac{1}{1+e^{-\langle a,w\rangle}} \tag{1}$$

1.2.1 输入

本文从宽度为 5 的上下文窗口中抽取特征。其中包括一元特征和二元特征，如表 1 所示。已有的研究表明，5 字长的上下文窗口恰好大致表达了前后各一个词的上下文[16]，从这个意义来讲，5 字宽的上下文窗口具备了字和词的双重含义，足以覆盖真实文本中绝大多数的构词情形。

表 1　一元/二元特征表

Table 1　Uni-and bi-gram feature template

特征类型	特征
一元特征	$C_{i-2},C_{i-1},C_i,C_{i+1},C_{i+2}$
二元特征	$C_{i-2}C_{i-1},C_{i-1}C_i,C_iC_{i+1},C_{i+1}C_{i+2}$

其中，下标代表了字与待标注字的相对位置。C_i 表示当前待标注字，C_{i-1} 表示当前待标注字的前一个字，C_{i+1} 表示当前待标注字的后一个字，依此类推。本文以上例中的“睦”“谐”分别作为当前待标注字，则“睦”对应的一元特征分别为“定”“和”“睦”“的”“两”，对应的二元特征分别为“定和”“和睦”“睦的”“的两”；同理，“谐”对应的一元特征分别为“主”“和”“谐”“的”“气”，对应的二元特征分别为“主和”“和谐”“谐的”“的气”。

1.2.2 输出

字符序列中的每一个字符都有确切的词位标注结果。本文使用“S”和“C”两种标签表示当前待标注字可能的标记状态。其中“S”(Separation)标签表示当前字与前一个字处于分离状态，即以当前字符开始一个新的词，而“C”(Combination)标签表示当前字与前一个字处于结合状态，即与前一个字组成一个词或词的一部分。以下面的句子为例，其正确标记序列如下。

建-S 立-C 稳-S 定-C 和-S 睦-C 的-S 两-S 岸-C 关-S 系-C。-S

本文采用的激活函数输出值分布在(0,1)内。本文以 0.5 作为阈值，若输出值大于 0.5 则标记为“S”，否则标记为“C”。

1.3 模型训练

本文采用交叉熵作为损失函数，如式(2)所示。训练过程首先在当前参数下预测待标注字的标记，再根据语料中其正确标记来更新模型参数。

$$J=-\sum_{i=1}^{n} y_i \log(\mathrm{h}(a_i))+(1-y_i)\log(1-\mathrm{h}(a_i)) \tag{2}$$

其中：y_i 表示正确标记状态，$\mathrm{h}(a_i)$ 为模型预测的结果。

为了防止过拟合导致模型的泛化能力降低，本文在损失函数中增加 ℓ_2 正则项，如式(3)所示。其中 λ 为正则项系数，用于控制正则化的强度。

$$J=J+\frac{\lambda}{2}(\|a\|^2+\|w\|^2) \tag{3}$$

本文采用随机梯度下降法对目标函数进行优化，采用误差反向传播的方式分别求出目标函数对 w 和 a 的梯度，更新 w 时保证 a 不变，反之亦然。更新公式如式(4）(5)所示。

$$a=a-\eta\frac{\partial J}{\partial a} \tag{4}$$

$$w=w-\eta\frac{\partial J}{\partial w} \tag{5}$$

1.4 基于聚类的构词规律分布结构发现

在无监督学习中聚类算法可以用于寻找数据内在的分布结构。本文以层次聚类[17]作为后期模型训练的前驱工作，用于发现基于字的多模型分词方法中模型参数所表示的构词规律分布结构。层次聚类首先将每个模型参数作为一个类别，然后根据距离不断合并这些原子类,形成一个具有树型的聚类结构，最后根据事先设定的簇间切分标准对聚类结构进行切分，形成最终的类簇。具体算法流程如下：

a)将每个模型参数看做一类，计算两两之间的距离；

b)将距离最小的两个类合并成一个新类；

c)重新计算新类与所有类之间的距离；

d)重复(2)(3)，生成一个具有树型的聚类结构；

e)根据簇间切分标准对聚类结构进行切分，形成最终的类簇。

聚类结束后，将得到字的类簇。字所处类簇相同，代表该类字的构词规律相同或相近;反之，则说明字的构词规律存在差异。

层次聚类算法中通常采用的距离度量方式为欧氏距离、余弦相似度，具体如式(6)(7)所示。

$$dist_{ed}=\sqrt{\sum_{n=1}^{d}|x_n-x'_n|^2} \tag{6}$$

$$dist_{\cos}=\frac{\sum_{n=1}^{d}x_n x'_n}{\sqrt{\sum_{n=1}^{d}x_n^2}\sqrt{\sum_{n=1}^{d}x'^2_n}} \tag{7}$$

其中：x, x' 代表模型参数，d 表示向量的维度，$dist_{ed}$ 代表欧氏距离，值越小，表示两个模型参数之间距离越小，即两个模型参数越相近；$dist_{\cos}$ 代表余弦相似度，值越大，表示两个模型参数之间夹角越小，两个模型参数距离越小，即两个模型参数之间越相近。

算法执行过程中为了统一采用距离最小值作为类簇合并条件，在使用余弦相似度作为距离度量方式时实际采用 $1\text{-}dist_{\cos}$，当 $1\text{-}dist_{\cos}$ 值越小，表示两个模型参数越相近。

簇间切分标准采用不一致系数，该系数反映了树型聚类结构中两个类簇合并时的距离与其下层深度为 2 的类簇合并时的距离不一致程度。当划分到有明显区别的类簇时，不一致系数较高，反之亦然。不一致系数计算公式如式(8)所示。

$$inconsistency=\frac{h-avg}{std} \tag{8}$$

其中：$inconsistency$ 代表不一致系数，h 代表合并的两个类簇的距离，avg 表示下层深度为 2 的类簇合并时距离平均值，std 表示下层深度为 2 的类簇合并时距离标准差。当两个类簇的不一致系数小于阈值时，将这两个类簇合并为一个新类簇，否则，将两个类簇进行切分。

1.5 分词算法的性能分析

分词算法性能的优劣往往通过算法复杂度来衡量。其中包括时间复杂度和空间复杂度。机器学习中，由于分词训练过程为一次性行为，分词时并不需要重新训练模型，只需将已训练好的模型用于分词任务。因此，对模型训练所消耗的时间代价关注相对较低，在实际的分词过程中更多的关注分词速度，以及模型存储所占的存储空间。

本节中将重点分析分词过程的时间复杂度。分词过程需要查找模型训练时生成的 Look-up 表,以及模型参数 **w**。已知需要进行分词的字数为 m，Look-up 表以及模型参数的数量为 n。查找 Look-up 表和模型参数的时间复杂度均为 O(1)。分词时每对一个字进行词位标注时均需要遍历整个 Look-up 表以及模型参数 **w**。因此，本方法中分词过程的时间复杂度为 O(mn) 。

与基于字的多模型分词方法相比，本文所提方法的模型数量远远小于基于字的多模型分词方法的模型数量，因此，在时间复杂度与空间复杂度方面均有优势。

2 实验及结果分析

2.1 数据和预处理

本文实验所采用的语料为 PKU 语料和 MSR 语料，它们是由 SIGHAN 举办的第二届国际中文分词评测 Bakeoff 2005 所提供的封闭语料。其中包括训练集、测试集、测试集的标准答案、词典以及评分脚本。其语料详细信息如表 2 所示。

表 2　PKU 语料和 MSR 语料的详细信息

Table 2　Corpus details of PKU and MSR

	PKU 语料	MSR 语料
词型	5.5×10^4	8.8×10^4
词例	1.1×10^6	2.4×10^6
字型	5×10^3	5×10^3
字例	1.8×10^6	4.1×10^6

实验中，在当前待标注字窗口宽度为 5 的上下文环境中提取特征时，当待标注字为句子开头或结尾字符时，需要在字符的左边(或右边)添加两个诸如“start-1”“start-2”(或“end-1”“end-2”)的字符进行补充，且补充字符的标记状态均为“S”。

2.2　评价方法

中文分词性能的评价指标通常采用准确率（P）、召回率（R）、F 值（F）、未登录词召回率（R_{OOV}）、词表词召回率（R_{IV}）。具体定义如下：

$$P=\frac{\text{系统正确识别的词语总数}}{\text{系统识别的词语总数}}\times 100\%$$

$$R=\frac{\text{系统正确识别的词语总数}}{\text{测试语料中的词语总数}}\times 100\%$$

$$F=\frac{2\times P\times R}{P+R}$$

其中：F 值作为分词性能评估的主要参考指标；未登录词召回率能够很好的反映模型的泛化能力。

2.3　实验参数设置

神经网络模型中超参数的选取对分词性能起着显著的影响，模型中各项超参数设置如表 3 所示。

表 3　神经网络模型中超参数设置

Table 3　Setting of the hyper-parameters

参数名称	值
学习率	eta = 0.01
迭代次数	iter = 100
损失容忍度	tol = 0.001
ℓ_2 正则项系数	ℓ_2 = 1
特征 Embedding 维度	d = 50

文献[1]中通过大量实验证明，特征 Embedding 的维度设定为 50 维，既能保证模型的训练速度又可以保证分词性能。因此，本文实验中将特征 Embedding 设置成 50 维。此外，为了防止模型训练过程中因某一错误导致训练无法终止，本文设置两种结束条件，一种为迭代次数，另一种为损失容忍度。当模型训练过程中满足两种情况中的一种，训练过程将停止。

2.4　实验结果及分析

为了验证本文所提方法的有效性，本文列举了如下几种分词方法进行对比。其中包括 CRF 分词方法、单模型分词方法、基于字的多模型分词方法以及神经网络分词方法。其中，单模型分词方法是指针对训练语料建立单一模型参数；基于字的多模型分词方法针对每个字建立单独的模型参数；CRF 分词方法采用 3 种策略，分别为 2 标记和 4 标记并考虑标记二元转移特征，记为 CRF2 和 CRF4,以及采用 2 标记方法但不考虑标记二元转移特征，记为 CRF。其中，CRF 分词实验均采用表 1 所示的特征模板；神经网络模型中特征 Embedding 均为随机初始化，进行了如下对比实验。

2.4.1 聚类算法距离度量方式及阈值选择实验

对基于字的多模型分词方法训练得到的字模型进行层次聚类，目的是将具有相同或相近构词规律的字聚合为一类，形成字的类簇。因此，得到聚类结果的好坏直接影响后续模型的再训练。表 4、5 展示了使用不同距离度量方式得到的实验结果。

表 4　在 PKU 语料中使用不同距离度量方式的实验结果

Table 4　Performances of using different distance metrics in PKU test set

距离度量方式	PKU 语料				
	P	R	F	R_{OOV}	R_{IV}
欧式距离	93.8	92.6	93.2	65.8	94.2
余弦相似度	**94**	**92.9**	**93.5**	**66.4**	**94.5**

从表 4、表 5 中可以看出，在两种语料中使用余弦相似度作为距离度量方式表现出最佳的分词结果。余弦相似度计算的为两个向量夹角的大小，反映两个向量之间的相似程度，而欧式距离则是度量两个向量之间位置的绝对距离。本文的聚类对象为模型参数，该向量反映了字的上下文特征对该字标注状态的作用程度，代表了该字的构词规律，使用余弦相似度作为距离度量方式更为合理。表 6 展示了使用余弦相似度作为距离度量方式下的聚类结果。

表 5　在 MSR 语料中使用不同距离度量方式的实验结果

Table 5　Performances of using different distance metrics in MSR test set

距离度量方式	MSR 语料				
	P	R	F	R_{OOV}	R_{IV}
欧式距离	95.5	95.5	95.5	53.9	**96.6**
余弦相似度	**95.6**	**95.6**	**95.6**	**55.9**	**96.6**

表 6　与“吴”“鸯”“扒”“蚣”相近的字

Table 6　Characters similar to “吴”“鸯”“扒”“蚣”

字	距离度量方式	构词规律相近的字
吴	余弦相似度	赵、彭、徐、卢、蔡
鸯		鸠、鹋、鹃、鹊、鸪
扒		扔、扛、扭、抛、抡
蚣		蟀、蜓、蟾、蛾、螂

如表 6 所示，“吴”为姓氏，通过聚类得到与之相近的几个字中“赵”“彭”“徐”“卢”“蔡”也作为姓氏，具有相同的构词规律。同理，“鸯”、“扒”、“蚣”分别作为鸟类名称、动词、昆虫名称，分别得到与之对应的类别字组成的字簇。说明对基于字的多模型分词方法学习出的模型参数进行聚类分析可以有效地获得相同或相近构词规律的字簇。

与 K-Means 聚类方法有所不同，层次聚类无须事先指出具体的聚类个数，而是通过设定不一致系数阈值得到最优的类簇，因此，不一致系数阈值的设定对实验结果有一定的影响。图 3 展示了使用不同的不一致系数得到的实验结果。

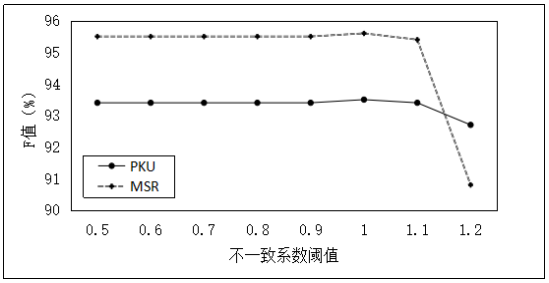

图 3　不一致系数阈值对分词性能影响

Fig. 3　Performance of using different inconsistentcy

从图 3 中可以看出，在两种语料中，使用层次聚类算法对基于字的多模型分词方法训练得到的模型参数进行聚类分析，不一致系数阈值设定为 1 时，得到最优的聚类结果，分词效果最佳。因此，在接下来的分词实验中将继续使用不一致系数阈值为 1 的这一设置。

2.4.2 模型对比实验

表 7、8 给出了本文所提出方法与单模型方法、基于字的多模型方法实验对比结果。从中可以看出，在两种语料上，本文所提出的方法表现出较优越的分词性能。其中，在 PKU 语料上，F 值高于单模型 1.2 个百分点，高于多模型 0.1 个百分点；在 MSR 语料上，F 值高于单模型 4.4 个百分点，高于多模型 0.1 个百分点，表现出足够的稳定性。单模型方法在未登录词识别方面表现出明显的优势，考虑上下文环境对所有字的全局综合作用，学习出字构词的一般性规律；而多模型针对每个字进行建模，学习出字构词的特殊性规律，因而在词表词召回率方面表现出强有力的优势。本文方法通过对模型参数聚类，将两种建模思想进行结合，即学习出一般性

构词规律又学习出特殊性构词规律，则分词效果优于其他两种方法。

表 7　在 PKU 语料上的实验结果对比

Table 7　Comparison with performance on PKU corpus

模型	PKU 语料				
	P	R	F	R_{OOV}	R_{IV}
单模型	93.3	91.7	92.3	**71.2**	93.0
多模型	93.9	92.8	93.4	65.6	**94.5**
this approach	**94**	**92.9**	**93.5**	66.4	**94.5**

表 8　在 MSR 语料上的实验结果对比

Table 8　Comparison with performance on MSR corpus

模型	MSR 语料				
	P	R	F	R_{OOV}	R_{IV}
单模型	91.8	90.6	91.2	**60**	91.4
多模型	95.4	95.5	95.5	53.8	**96.6**
this approach	**95.6**	**95.6**	**95.6**	55.9	**96.6**

与此同时，本文对比了上述三种分词方法的模型数量，结果展示在表 9 中。与多模型相比，在 PKU 语料上，模型数量由 4686 减少到 1854，减少幅度近五分之三；在 MSR 语料上，模型数量由原来的 5151 减少到 2299，减少幅度近二分之一。说明本文的实验方法在提高 F 值的同时，大幅度减少模型数量，节约了模型存储成本。

表 9　在两种语料上模型数量对比结果

Table 9　Comparison with model numbers on two corpora

模型	PKU 语料		MSR 语料	
	F 值	模型数	F 值	模型数
单模型	92.3	1	91.2	1
多模型	93.4	4686	95.5	5151
this approach	93.5	1854	95.6	2299

为了进一步验证本文所提方法的有用性，将本文所提分词方法与单模型分词方法、多模型分词方法在分词时间与模型存储所占空间方面进行对比。其中分词时间是指利用已训练好的模型进行分词时所消耗的时间；存储空间则是指模型存储所占空间的大小。实验结果展示在表 10 中。

表 10　两种语料上分词时间与模型存储空间比较

Table 10　Comparison with word segmentation time and model storage space on two corpora

模型	PKU 语料			MSR 语料		
	F	分词时间(ms)	存储空间	F	分词时间(ms)	存储空间
单模型	92.3	536	4.3 KB	91.2	897	4.4 KB
多模型	93.4	1117	20 MB	95.5	1314	22 MB
this approach	93.5	1084	7.9 MB	95.6	1207	9.8 MB

从表 10 中可以看出，与多模型分词方法相比，本文所提方法在提高分词性能的同时，在分词时间与模型存储空间方面也具有一定的优势，尤其在模型存储空间方面，大幅度节约了存储成本，更有利于工程中实际的分词应用。分析原因，单模型基于训练语料建立单一模型参数，因此所占存储空间最少，分词速度最快；而基于字的多模型分词方法基于每个字进行建模，存在模型的冗余，模型所占存储空间较多；本文所提方法将具有相同或相近构词规律的字合并为字簇，进行模型训练，大大减少了模型数量，与多模型相比，提高分词性能的同时降低了模型存储所占空间。

表 11、12 比较了该方法与 CRF 分词方法的实验对比结果，可以看出，CRF 采用 4 标记并加入标记转移特征的模型表现出较好的分词性能。与本文所提出的方法进行对比，在 PKU 语料上，本文的方法在 5 种评价指标中皆高于 CRF4 方法，其中，F 值提升 0.4 个百分点；但在 MSR 语料上，CRF4 方法分词性能明显优于本文所提出的方法，F 值高出 0.8 个百分点。从 MSR 语料实验结果中可以看出，CRF4 方法与本文方法在词表词召回率方面相差不大，但在未登录词召回率上，CRF4 方法明显高于本文方法，导致 CRF4 方法的最终结果优于本文方法。分析原因，相比 PKU 语料，MSR 语料规模相对较大，CRF4 方法训练得更充分，对未登录词识别能力更强。

表 11　在 PKU 语料上与 CRF 实验对比结果

Table 11　Comparison with results used CRF on PKU corpus

模型	PKU 语料				
	P	R	F	R_{OOV}	R_{IV}
CRF	93.3	90.9	92.1	53.0	93.2
CRF2	93.1	91.2	92.1	55.6	93.3
CRF4	94.0	92.2	93.1	61.2	94.1
This approach	**94.0**	**92.9**	**93.5**	**66.4**	**94.5**

表 12　在 MSR 语料上与 CRF 实验对比结果

Table 12　Comparison with results used CRF on MSR corpus

模型	MSR 语料				
	P	R	F	R_{OOV}	R_{IV}
CRF	94.9	95.4	95.1	51.0	96.6
CRF2	95.4	95.1	95.3	63.6	95.9
CRF4	**96.5**	**96.2**	**96.4**	**70.8**	**96.9**
this approach	95.6	95.6	95.6	55.9	96.6

表 13　与前人工作进行对比

Table 13　Comparision with previous models

模型	PKU 语料			MSR 语料		
	P	R	F	P	R	F
Zheng et al(2013)	92.8	92.0	92.4	92.9	93.6	93.3
Pei et al.(2014)	93.7	93.4	93.5	94.6	94.2	94.4
Chen et al.(2015)	**95.8**	**95.5**	**95.7**	**96.7**	96.2	96.4
Cai et al.(2016)	95.5	94.9	95.2	96.1	**96.7**	96.4
this approach	94.0	92.9	93.5	95.6	95.6	95.6

本文将实验结果与相同数据集上的前人工作进行了对比。如 2013 年，Zheng 等人[18]应用 Collobert 等人[19]的神经网络框架进行分词；2014 年，Pei 等人[20]通过利用标签嵌入和基于张量的转换，提出了 MMTNN 的神经网络进行分词；2015 年，Chen 等人[21]为了解决中文分词中无法长期依赖信息的问题，提出了 LSTM 神经网络并用于分词；2016 年，Cai 等人[22]利用门控组合神经网络对字符进行分布式表示，并利用 LSTM 神经网络对预测结果进行打分。实验结果对比如表 13 所示。与 Zheng 等人相比，本文所提方法在 PKU 语料上 F 值提高 1.1 个百分点，在 MSR 语料上 F 值提高 2.3 个百分点；与 Pei 等人相比，本文所提方法在 PKU 语料上 F 值达到了相一致，在 MSR 语料上 F 值高出 1.2 个百分点。与 Chen 等人和 Cai 等人实验结果相比，本文方法的分词性能略显不足。

将本文分词结果与 Cai 等人的分词结果作进一步对比分析，发现本文所提方法在切分诸如“入/军队”、“服/现役”、“战/风雪”、“拟/任”、“求/发展”等单字动词时的切分效果优于 Cai 等人的分词方法。此外，对本实验中具体的分词结果进行分析发现如表 14 中所列的切分错误。

表 14　切分结果对比

Table 14　Comparison with segmentation results

正确切分方式	错误切分方式
就业/旺季	就业旺季
无/党派/人士	无党派人士
浙江/海盐县	浙江海盐县
世纪/交替	世纪交替

从上述例子中可以看出，切分结果出现多词粘连的情况。该种切分错误在文献[23]中均有谈到，该文献中通过实验验证了基于字的分词方法往往忽略词所包含的组合信息，指出应用字词联合解码进行分词效果更佳。通过分析本文切分错误结果，同样验证了上述结论，而本文所提出的分词方法，恰好缺少词信息进行分词指导学习，因此出现多词粘连的情况，影响了最终的分词性能。对比 Cai 等人的分词方法，通过门控组合神经网络对输入的字符序列进行候选词分布式表示，很好地引入了词信息，并用 LSTM 神经网络对所有切分结果进行打分，取打分最高的切分组合作为最终的分词结果，则最终的分词效果优于本文实验结果。在今后的实验中，本文将借鉴 Cai 等人的分词方法，引入词信息进行指导学习。

3　结束语

本文提出了一种基于字簇的多模型中文分词方法，该方法可以看做单模型与基于字的多模型建模思想的结合，很好地发挥了单模型分词方法发现未登录的作用以及基于字的多模型分词方法切分词表词的作用，学习出字构词的一般性与特殊性构词规律。实验结果表明，与基于字的多模型分词方法相比，该方法在小幅度提升分词性能的同时，有效减少了模型数量，降低了模型存储成本并且提升了分词速度。

通过实验部分分析，本文的分词方法并没有引入词的信息对分词过程进行指导学习，影响了最终的分词性能，有一定的局限性。今后的工作中，可以尝试加入词的信息，提高分词质量；另一方面，本文的方法是利用聚类发现字构词之间分布结构规律，聚类的好坏直接影响分词的效果，今后的工作中可以在算法层面做进一步的尝试，利用多任务学习算法进行分词实验[24-26]。

参考文献：

[1] 来斯惟，徐立恒，陈玉博，等. 基于表示学习的中文分词算法探索 [J]. 中文信息学报, 2013, 27(5): 8-14. (Lai Siwei, Xu Liheng, Chen Yubo, *et al.* Exploring Chinese word segmentation algorithm based on representation learning [J]. Journal of Chinese Information Processing, 2013, 27(5): 8-14.)

[2] Xue Neiwen, Shen Libin. Chinese word segmentation as LMR tagging [C]//Proc of the 2nd SIGHAN Workshop on Chinese Language Processing. New York: ACM Press, 2003: 176-179.

[3] McCallum A, Freitag D, Pereira F. Maximum entropy markov models for information extraction and segmentation [C]//Proc of International Conference on Machine Learning. New York: ACM Press, 2000: 591-598.

[4] 李月伦，常宝宝. 基于最大间隔马尔可夫网模型的汉语分词方法 [J]. 中文信息学报, 2010, 24(1): 8-14. (Li Yuelun, Chang Baobao. Chinese word segmentation method based on maximum interval markov network model [J]. Journal of Chinese Information Processing, 2010, 24 (1): 8-14.)

[5] Tseng H, Chang P, Andrew G, *et al.* A conditional random field word segmenter for sighan bakeoff 2005 [C]// Proc of the 4th SIGHAN Workshop on Chinese Language Processing. New York: ACM press, 2005: 168-171.

[6] Zhang Ruiqiang, Kikui G, Sumita E. Subword-based tagging by conditional random fields for Chinese word segmentation [C]// Proc of Human Language Technology Conference of the North American Chapter of the ACL. Stroudsbury, PA: ACL, 2006: 193-196.

[7] Zhao Hai, Huang Changning, *et al.* Effective tag set selection in Chinese word segmentation via conditional random field modeling [C]// Proc of Pacific Asia Conference on Language, Information and Computation. New York: ACM press, 2006: 87-94.

[8] He Jia, Li Guanghong. Research of Chinese word segmentation based on neural network and particle swarm optimization [C]// Proc of the 3th International Conference on Apperceiving Computing and Intelligence Analysis. Piscataway,NJ: IEEE Press, 2010: 56-59.

[9] Zheng Xiaoqing, Chen Hanyang, Xu Tianyu. Deep learning for Chinese word segmentation and POS tagging [C]// Proc of the 18th Conference on Empirical Methods in Natural Language Processing. Stroudsbury, PA: ACL, 2013: 647-657

[10] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, et al. Gated recursive neural network for Chinese word segmentation [C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsbury, PA: ACL, 2015: 1744-1753

[11] Cai Deng, Zhao Hai, Zhang Zhisong, *et al.* Fast and accurate neural word segmentation for Chinese [C]// Proc of the 55th Annual Meeting of Association for Computational Linguistics. Stroudsbury, PA: ACL, 2017: 608-615.

[12] 韩冬煦，常宝宝. 中文分词模型的领域适应性方法 [J]. 计算机学报, 2015, 38(2): 272-281. (Han Dongxu, Chang Baobao. Domain adaptation method of Chinese word segmentation model [J]. Chinese Journal of Computers, 2015, 38(2): 272-281.)

[13] Qiu. Likun, Zhang Yue. Word segmentation for Chinese novels [C]// Proc of the 29th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 2440-2446.

[14] 张少阳，王裴岩，蔡东风. 一种基于字的多模型中文分词方法 [J]. 沈阳航空航天大学学报, 2017, 34 (1): 70-75. (Zhang Shaoyang, Wang Peiyan, Cai Dongfeng. A multi-model of Chinese word segmentation based on character [J]. Journal of Shenyang Aerospace University, 2017, 34 (1): 70-75.)

[15] Ma Jianqiang, Hinrichs E. Accurate linear-time Chinese word segmentation via embedding matching [C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsbury, PA: ACL, 2015: 1733-1743

[16] 黄昌宁，赵海. 中文分词十年回顾 [J]. 中文信息学报, 2007, 21(3): 8-19 (Hang Changning, Zhao Hai. Ten years of Chinese participle review [J]. Journal of Chinese Information Processing, 2007, 21(3) 8-19.)

[17] 孙吉贵，刘杰，赵连宇. 聚类算法研究 [J]. 软件学报, 2008, 19(1): 48-61. (Sun Jigui, Liu Jie, Zhao Lianyu. Clustering algorithm research [J]. Journal of Software, 2008, 19 (1): 48-61.)

[18] Zheng Xiaoqing, Chen Hanyang, Xu Tianyu. Deep learning for Chinese word segmentation and POS tagging [C]// Proc of the 18th Conference on Empirical Methods in Natural Language Processing. Stroudsbury, PA: ACL, 2013: 647-657

[19] Collobert R, Weston J, Bottou L, *et al.* Natural language processing

(almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12 (1): 2493-2537

[20] Pei Wenzhe, Ge Tao, Chang Baobao. Max-margin tensor neural network for Chinese word segmentation [C]// Proc of the 52th Annual Meeting of the Association for Computational Linguistics. Stroudsbury, PA: ACL, 2014: 293-303.

[21] Chen Xinchi, Qiu Xipeng, Zhu Chenxi, *et al*. Long short-term memory neural networks for Chinese word segmentation [C]// Proc of the 20th Conference on Empirical Methods in Natural Language Processing. Stroudsbury, PA: ACL, 2015: 1197-1206.

[22] Cai Deng, Zhao Hai. Neural word segmentation learning for Chinese [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsbury, PA: ACL, 2016: 409-420.

[23] 宋彦，蔡东风，张桂平，等．一种基于字词联合解码的中文分词方法 [J]. 软件学报，2009, 20(9): 2366-2375. (Song Yan, Cai Dongfeng, Zhang Guiping, *et al*. A Chinese word segmentation method based on joint decoding of words [J]. Journal of Software, 2009, 20(9): 2366-2375.)

[24] Liu Jun, Ji Shuwang, Ye Jieping. Multi-task feature learning via efficient $L_{2,1}$-norm minimization [C]//Proc of Conference on Uncertainty in Artificial Intelligence. 2009: 339-348.

[25] Chen Xinchi, Shi Zhan, Qiu Xipeng, *et al*. Adversarial multi-criteria learning for Chinese word segmentation [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsbury, PA: ACL, 2017: 1193-1203.

[26] Liu Pengfei, Qiu Xipeng, Huang Xuanjing. Adversarial multi-task learning for text classification [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsbury, PA: ACL, 2017: 1-10.